Staged Designs in Whole-genome Association Studies

Created 19/4/2006

Background

The idea of staged design is not new. However, its objective largely remains unchanged, i.e., to reduce cost without compromising efficiency. This becomes eminent recently in statistical genetics with the so-called whole-genome association (WGA) studies (Carlson et al. 2004, Hirschhorn & Daly 2005, Wang et al. 2005), which typically involve tens of thousands of single nucleotide polymorphisms (SNPs), the most common form of variants in human genome. They are believed to hold the key to common diseases and population history; several recent papers provided such evidence (e.g. Klein et al. 2005, Herbert et al. 2006, Grant et al. 2006). Here I present some materials that I came across during the past few months but expose one paper in an effort to keep up with the latest literature. The relevance to several studies carried out in the Unit is also described. A summary of an informal discussion within the DiOGenes project is later attached.

The early work on staged design was in line with the development in genetic epidemiology in general, e.g. linkage studies by Elston et al. (1996), Holmans & Craddock (1997), Sham & Zhao (2000), Guo & Elston (2001), followed by association studies involving WGA by Satagopan et al. (2002, 2004), Satagopan & Elston (2003), Thomas et al. (2004), Skol et al. (2006), Lin (2006), Wang et al. (2006).

Issues

Given our study sample and SNPs of interest are defined, a staged design furnishes collection of all information in several steps. In the simplest and well-studied two-staged design of genetic case-controls studies, a proportion of individuals is genotyped at all of the SNPs and a proportion of the most significant ones is selected and to be carried over as replication study at the second stage.

The question related to statistics or operations research therefore regards how do we allocate our resources between stages? by which criteria? The optimal design depends upon the purpose of the analysis (e.g. estimation or hypothesis testing, estimating relative risks or localizing a causal variant) and specific statistical methods to be used (e.g. hyplotype- or genotype-based) (Thomas et al. 2004). In general, one has to assume linkage equilibrium or disequilibrium, case-control or other types of designs, re-use of the genotype data on many phenotypes, gene-environment interaction, etc.

An aspect that was missed in most work is the analysis. The paper by Skol et al. (2006) made an attempt to address this. The main results can be seen through Figures 1 to 4 and Table 1 and a single take home message is that joint analysis of two-stage data is more powerful (discussions on Figures 1-4 and Table 1).

Notations

Among others, the following parameters are considered,

M = number of SNPs
N = number of cases/controls, or n_1 and n_2 if assuming unequal numbers
$\pi_{samples}$ = proportion of samples to use at stage 1
$\pi_{markers}$ = proportion of markers to be carried out at stage 2
α_{genome} = the genome-wide false positive rate
$\alpha_{markers}$ = the false positive rate when using stage 1 and stage 2
C = significance threshold for single stage involving all individuals
C_1 = significance threshold for stage 1
C_2 = significance threshold for stage 2
C_{joint}=significance threshold for joint analysis of stages 1 and 2 thresholds
$\hat{p}_1$ = the estimated risk allele frequency in cases
$\hat{p}_2$ = the estimated risk allele frequency in controls
p' = expected frequency of cases
p = expected frequency of controls
a = the observed staged 1 statistic
K = the population prevalence of disease
γ = genotype relative risk (GRR)

1. The statistic for case-control data

In a complete genotyping approach, the statistic formed by cases and controls is as follows,

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{2n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{2n_2}}}$$

The factor 2 in the denominator is because each individual contributes two copies of alleles. Under the null hypothesis it is well-known the statistic has a standard normal distribution, N(0,1). The power can be obtained at a given significance level, based on the significance threshold C and cumulative normal distribution $\Phi(.)$ (similar argument using chi-squared statistic was discussed earlier).

The statistic becomes,

$$z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{2n_1\pi_{samples}} + \frac{\hat{p}_2(1-\hat{p}_2)}{2n_2\pi_{samples}}}}$$

if only $\pi_{samples}$ of individuals is used at stage one.

As we have seen the expressions given above have similar form, for implementation purpose we can define a function as z(p1,p2,n1,n2,pi.samples). And z(p1,p2,n1,n2,1) would give the first expression while z(p1,p2,N,N,pi.samples) corresponds to expression in the paper.

2. Two-stage design

The usual two-stage design proceeds in the same manner by selecting only $\pi_{mar\,ker\,s}$ of the markers to be genotyped at stage two, leading to similar statistic but with $1-\pi_{mar\,ker\,s}$ in the denominator.

In order to compare power, the means and variances of these statistics under both the null and the alternative hypotheses will be sufficient. For large sample we are more comfortable to think the mean as "noncentral" parameter or displacement, denoted as μ_1

$$\mu_1 = \frac{p'-p}{\sqrt{\frac{p'(1-p')}{2n_1\pi_{samples}} + \frac{p(1-p)}{2n_2\pi_{samples}}}}$$

or in our implementation z(pprime,p,n1,n2,pi.samples). Again μ_2 is similarly obtained as z(pprime,p,n1,n2,1-pi.samples).

Now the probability that a marker remain significant at both stages one and two is $\alpha_{mar\,ker} = P(|z_1| > C_1)P(|z_2| > C_2, sign(z_1) = sign(z_2))$ where sign(.)=-1,0,1 according to the sign of the function's argument, or

$$P_2 = \frac{[1-\Phi(C_2-\mu_2)][1-\Phi(C_1-\mu_1)]}{1-\Phi(C_1-\mu_1)+1-\Phi(-C_1-\mu_1)} + \frac{\Phi(-C_2-\mu_2)\Phi(-C_1-\mu_1)}{1-\Phi(C_1-\mu_1)+1-\Phi(-C_1-\mu_1)}$$

The denominator is recognised as the power for stage 1, so the power for replication-based study is

$$P_1P_2 = [1-\Phi(C_2-\mu_2)][1-\Phi(C_1-\mu_1)] + \Phi(-C_2-\mu_2)\Phi(-C_1-\mu_1)$$

3. The statistic for joint analysis

For the joint analysis, the parameter of interest is as follows,

$$\mu_{joint} = \frac{p'-p}{\sqrt{\frac{p'(1-p')}{2n_1} + \frac{p(1-p)}{2n_2}}} + \sqrt{\pi_{samples}}\,a$$

Note the expression given in the paper is always $\sqrt{\pi_{samples}}\,a$; the correct way of achieving the required quantity under the null hypothesis is the 1st part is zero!

Getting this right, the power for the joint analysis is $P_1 P_{jo\,\mathrm{int}}$ with

$$P_{jo\,\mathrm{int}} = \int_{-\infty}^{-C_1} P(\left|z_{jo\,\mathrm{int}}\right| > C_{jo\,\mathrm{int}} \mid T = x)\, f(x \mid T)dx + \int_{C_1}^{\infty} P(\left|z_{jo\,\mathrm{int}}\right| > C_{jo\,\mathrm{int}} \mid T = x)\, f(x \mid T)dx$$

where $f(x \mid T) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z_1-\mu_1)^2}{2}}$

since z_1 is not observed and an integration is invoked. This is essentially an extension of the expression for replication study since we can then factor the integrand at each term. The mean and variance formulae are reminiscent of the standard result of bivariate normal distribution, BVN(0,0,1,1, ρ), when $\rho = \sqrt{\pi_{samples}}$, e.g. Morton et al. (1983), p50. The conditional mean is ρa and variance $1-\rho^2$.

4. Final notes

4.1 Significance levels

The paper sets p values at stage 1 less than $\pi_{mar\,\mathrm{ker}\,s}$, and $\alpha_{mar\,\mathrm{ker}} / \pi_{mar\,\mathrm{ker}\,s}$ for replicate analysis. For a joint analysis, however, they use $\alpha_{mar\,\mathrm{ker}}/(M\pi_{mar\,\mathrm{ker}\,s})$ which appears to be a typo.

4.2 The efficiency of stage design

Table 1 (the 3rd column) contains proportion of genotypes, can be obtained through procedures just described with comparable power.

4.3 The estimated versus expected frequencies

The distinction between estimated versus expected allele frequencies is not obvious. Again we write out the following equation (see also Table 1).

$$K = p_1^2 f_2 + 2p_1(1-p_1) f_1 + (1-p_1)^2 f_0$$

Table 1. The penetrance function under four disease models

	The penetrance function		
Model	f_0	f_1	f_2
Multiplicative	1	γ	γ^2
Additive	1	γ	$2\gamma-1$
Recessive	1	1	γ
Dominant	1	γ	γ

To make these penetrances compromise with the prevalence, they requires to be rescaled. The expected frequencies of cases and controls are then,

$$p' = \frac{f_2 p_1^2 + f_1 p_1 (1 - p_1)}{K}$$ and

$$p = \frac{(1 - f_2) p_1^2 + (1 - f_1) p_1 (1 - p_1)}{1 - K}$$

respectively.

5. Summary

Given that there is a lot of work on staged designs, a "global" optimum is far from clear, and some mathematically less vigorous arguments are often required. In the summary by Thomas et al. (2005) they chose a less statistical word "sense".

Further to this point, a replication study aims to reduce false discovery rates, not to improve power as is the focus of Skol et al. (2006).

Briefly, the paper considers the statistic $z_{joint} = \sqrt{\pi_{samples}} z_1 + \sqrt{1 - \pi_{samples}} z_2$ and compares the appropriate significance thresholds under a variety of scenarios. The power is obtained as $P_1 P_2$ or $P_1 P_{joint}$ for replication or joint analysis using the usual expression $P(AB) = P(A)P(B \mid A)$, or
$P(|z_1| > C_1) P(|z_2| > C_2, sign(z_1) = sign(z_2))$ for replication and
$P(|z_1| > C_{joint}) P(|z_2| > C_{joint})$ for joint analysis.

The paper does present some results on proportion of genotypes (see Table 1), however. A recent work on cost efficiency is Wang et al. (2006).

A list of errors and typos

Table 1. $\pi_{samples}$ =0.20, columns C_2 and C_{joint} were incorrect.

Page 212, 2nd column, line -3 of the 2nd paragraph C_2=4.79 should be C_2=4.65.

Page 212, 2nd column, last equation should not contain μ_1.

Page 213, 1st column, line 4, $\alpha_{marker} / (M \pi_{samples})$ should be α_{genome} / M ?

Related work

With the EPIC 5000 design, it was proposed that 50% samples will be used at stage 1, with 10% markers selected at stage 2. This was in line with results on Table 1 and that by Lin (2006). However, this might still be suboptimal. It should be recognised that our data might not be standard in the sense that most individuals were already obese at baseline. Of course trait such as BMI is continuous. One drawback of the method by Skol et al. (2006) is its

assumption of linkage equilibrium which is not true given the number of SNPs in question, and simple Bonferroni correction is suboptimal. The work by Satagopan & Elston (2003) has similar limitation, and somewhat considered by Satagopan et al. (2004) and Thomas et al. (2004). Further aspects, including description of a three-stage design were described in Prentice et al. (2005). In the EPIC study of breast cancer, 13K of the ~ 250K SNPs at stage 1 were carried over to the second stage. An alternative argument uses only top 10~20 SNPs, described by Laird & Lange (2006) and Herbert et al. (2006).

There are other aspects to take further, namely application of the Genetic Analysis Workshop data (GAW 15) and the context of case-cohort design. Of course these are not necessarily limited to the literature outlined here.

References

Carlson et al (2004) Nature 429: 446-52

Elston et al. (1996) Genet Epidemiol 13:535-558

Grant et al. (2006) Nat Genet 38:320-323

Guo & Elston (2001) Advances in Genetics 459-471

Herbert et al. (2006) Science 312:279-283

Hirschhorn & Daly (2005) Nat Rev Genet 6:95-108

Holmans & Craddock (1997) Am J Hum Genet 60:657-666

Laird & Lange (2006) Nat Rev Genet 7:385-394

Morton et al. (1983) Methods in Genetic Epidemiology. Karger

Satagopan et al. (2002) Biometrics 58:163-170

Satagopan & Elston (2003) Genet Epidemiol 25: 149-157

Satagopan et al. (2004) Biometrics 60:589-97

Thomas et al. (2004) Genet Epidemiol 27: 401–414

Thomas et al. (2005) Am J Hum Genet. 77(3):337-45 (check closely with its supplemental information, and Table A2 formalised in Wang et al. (2006))

Prentice et al. (2005) Biometrics 61, 899–941 (read carefully on the comments by Thomas on genotype errors between stages originally reported by Clayton, see also the AJHG paper by Thomas)

Lin (2006) Am J Hum Genet 78:505-509 (a Fortran program/source is available)

Sham & Zhao (2000) GeneScreen 1:103-106

Skol et al. (2006) Nat Genet.38(2):209-13 (check the associate website for a program called CaTS as a toy for statisticians).

Wang et al. (2005) Nat Rev Genet 6: 109-18

Wang et al. (2006) Genet Epidemiol 30: 356–368

Appendix

Adaptation of some discussions on pros and cons of staged design in the DiOGenes project

These are from the e-mails to the DiOGenes list. It is not a complete collection but helps to formalise ideas if necessary.

Confusions on staged design (Thorkild Sorensen)

Should we go on analysing gene-environment interaction if there were no main effects, particularly for stage 1? Would we test other gene-environment interaction on the basis of the 1st stage data than those as our basic hypothesis? Should we trust gene-environment interaction findings that show up in the second but did not show up in the first round, either because of low power in the first phase or because of the heterogeneity in the interaction? Would we trust a significant interaction that emerged in the 1st round but disappeared in the second round, even though it might be due to cohort heterogeneity? Should we simply use the 2nd stage data but not joint analysis in order to reduce the risk of confusing results? What will we do if there is nothing to see after the first round?

Pros on staged designs (Thorkild Sorensen)

Notwithstanding the confirmatory testing of provisional apparently significant results obtained from a 1st phase is generally a good idea but is likely impractical. Given heterogeneous subcohorts involved, this will be a merit.

The two-stage hypothesis testing should be based on a random split of the entire cohort! *In other words, it is not that some cohorts are selected and genotyped for all SNPs and others ended up with few. This is somewhat contradictory to the claim above.*

It is possible to make a so-called conditional power estimation after the 1st phase: what is the probability of significance on the 2nd stage but no at all at the 1st stage?

Resources saved by staged design could be re-allocate for new genes.

Pros on two-stage design (Wim HM Saris)

At this moment we should leave the costs outside the discussion.

One important argument for the two-stage approach is the experience from NUGENOB such that not many SNPs could be linked to obesity-related issues. One argument is of course that the power was reduced, but one should suspect about the physiological relevance if not seen with 5000.

Pros on the single-stage design (Ruth Loos)

Two-stage designs are now recommended with no a priori hypotheses, no intention to fully characterise genes, intention to cover the (whole) genome as much as possible but expensive due to 10~500K SNPs. If one would not have to bother about funding, he/she would opt for genotyping all 10~500K SNPs in all available samples.

The full characterisation of each gene will allow analysis of complete haplotypes, without loss of any information.

In practice, a few SNPs broken up the haplotypes at stage 1 may end up in genotyping (almost) all SNPs again in the 2nd round, as the chance of finding a significant p-value in at least one SNP of each gene is high. The full characterisation is important since there is insufficient information available (i.e. HapMap). This is also the way to go in the future.

The advantage is that we don't need to re-design the chip which will give ample space for new candidate genes.

We hope to find interaction effects with clinical relevance, but those which are not clinically relevant can still teach us about the biology and underlying pathways.

Regardless of the design, the time schedule will be about the same.